DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2 Article 22

11-1-2013

Bayesian Joinpoint Regression Model for Childhood Brain Cancer Mortality

Ram C. Kafle
University of South Florida, Tampa, FL, rckafle@mail.usf.edu

Netra Khanal
The University of Tampa, Tampa, FL, nkhanal@ut.edu

Chris P. Tsokos
University of South Florida, Tampa, FL, ctsokos@usf.edu

Bayesian Joinpoint Regression Model for Childhood Brain Cancer Mortality

Erratum

In the original running head of this article, we incorrectly spelled author Netra Khanal's surname. We regret this error, and have corrected the article.

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 358-370.

ISSN 1538 − 9472

Bayesian Joinpoint Regression Model for Childhood Brain Cancer Mortality

Ram C. Kafle
University of South Florida
Tampa, FL

Netra Khanal
The University of Tampa
Tampa, FL

Chris. P. Tsokos
University of South Florida
Tampa, FL

The Bayesian approach of joinpoint regression is widely used to analyze trends in cancer mortality, incidence and survival data. The Bayesian joinpoint regression model was used to study the childhood brain cancer mortality rate and its average percentage change (APC) per year. Annual observed mortality counts of children ages 0-19 from 1969-2009 obtained from Surveillance Epidemiology and End Results (SEER) database of National Cancer Institute (NCI) were analyzed. It was assumed that death counts are probabilistically characterized by the Poisson distribution and they were modeled using log link function. Results were compared with the mortality trend obtained using joinpoint software of NCI.

Keywords: Bayesian statistics, brain cancer, joinpoint regression, mortality, SEER.

Introduction

The social and economic burden due to cancer is growing and is the major public health problem in the United States. Brain cancer (brain tumor and other central nervous system (CNS) cancers) is one of the leading cancers ranking the second largest cause of childhood death due to cancers. Based on 1975-2007 incidence data reported by Kohler, et al. (2011), 65.2 percent of the children with brain tumors are diagnosed with malignant tumor whereas the percentage in adult is only 33.7. According to National Cancer Institute (NCI), leukemias and the cancers of the brain and nervous system in children account for more than half of the new cases. Brain tumors are the most common solid tumors and are the second most common type of pediatric cancer. The central brain tumor registry of the United States reports that approximately 4,300 children younger than age 20 are expected to be diagnosed with primary malignant and non-malignant brain

Ram C. Kafle is a PhD candidate in Statistics. Email him at: rckafle@mail.usf.edu. Dr. Netra Khanal is an Assistant Professor of Mathematics. Email him at: nkhanal@ut.edu. Dr. Chris P. Tsokos is a Distinguished University Professor in Mathematics and Statistics. Email him at: ctsokos@usf.edu.

cancer in 2013. According to Kleihues, et al. (1993), the histological appearances of childhood brain tumors differ significantly from that of adult and are classified into several large groups. The overall distribution of these tumors also differ significantly (Peterson, et al., 2006; Pollack, 1994; Pollack, 1999). Ullrich and Pomeroy (2003) reported that the Pilocytic astrocytoma is the main histologic types in children CNS tumors with relatively high frequency of occurrence. According to Ries et al. (2007), the overall incidence for childhood brain cancer rose from 1975 to 2004 with the greatest increase occurring from 1983 through 1986. But, it is found that the mortality rates are continuously decreasing, with relatively higher rate from 1969 to 1980 and slower rate from 1980 onwards. These previous works provide motivation to study the mortality trend in childhood brain cancer using a statistical model that is based on realistic assumptions.

The main objective of this study is to give the reliable estimates of the measure of cancer mortality trend that provide up-to-date information and recent changes in childhood brain cancer. The joinpoint regression model is preferable when analyzing the trend for several years as it enables the identification points in the trend where the significant changes occur (Kohler, et al., 2011). If it is assumed that the joinpoints are random variables that can occur at any locations within the data range, the log likelihood is not differentiable with respect to break points suggesting that the Bayesian method is a more realistic approach. The actual Bayesian Joinpoint Regression Model will be solely based on Bayesian model selection criteria with the smallest number of joinpoints that accurately describe the Annual Percentage Change (APC) in the trend of mortality rates. Having good estimates of the mortality rates will allow the detection of points in time where significant changes occur and provide the best possible predictions. More practically, it helps to monitor the progress being made in childhood brain cancer, and evaluate the effectiveness of current treatment methods with respect to the mortality rate.

The history of joinpoint is not very long. In 1992, Charlin et al. developed hierarchical Bayesian analysis of changepoint problem in which they used an iterative Monte Carlo method. Kim et al. (2000, 2004) proposed a nonparametric approach which is widely used for analyzing and predicting the mortality and incidence data. NCI still uses this methodology, among others to find the trends in mortality, incidence, and survival of cancers in the United States. Tiwari et al. (2005) first developed a Bayesian model selection method for joinpoint regression. They discussed two criteria to select the best model, one with smallest BIC and other related to the Bayes factor. All of the previous studies assumed that the

errors are IID normal which is not always relevant with the real application data such as mortality and incidence of a specific disease in a population. This normality assumption is relaxed by Ghosh et al. (2009) proposing a Bayesian approach on parametric and semi-parametric joinpoint regression model. They introduced a continuous prior for the joinpoints induced by the Dirichlet distribution. The generalized linear model with log link function in joinpoint regression model that evaluates and incorporates the uncertainty in both model selection and model parameters has been recently introduced and implemented by Martinez-Beneito et al. (2011).

Studied here is the mortality trend of childhood brain cancer data obtained from SEER database of NCI. The total annual observed mortality counts of children below 20 years of age from 1969-2009 is extracted. Being rare events, assume the mortality counts are probabilistically characterized by the Poisson probability distribution and model them using log link function. The Bayesian joinpoint regression model discussed previously was used to obtain the mortality trend assuming that the break points are continuous over time. The joinpoint regression model is also fitted using the joinpoint software of NCI for the same data and compare these two results. Observe that the model using Bayesian approach describes the data very well giving best possible short term predictions and performs a better improvement over the existing methods.

Joint Point Model

Let $Y_i, i = 1, 2, ..., n$ be the number of mortality counts during a period of time t_i in a population. Let there be k change points that describe the behavior of the data, then the mean of the above outcome process can be expressed as the following generalized linear model

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \beta_j (t_i - \tau_j)^+, \quad (1)$$

where $\bar{t}$ is the mean of t_i, and τ_j is the change point in the model and g is monotonic and differentiable function, called the link function. The value of $(t_i - \tau_j)^+$ is $(t_i - \tau_j)$ if $(t_i - \tau_j)^+ > 0$ and 0 otherwise. For example, if there is no breakpoint in the model then

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t});$$

and if there is one break point, the model becomes

$$g\left[E\left(Y_i \mid t_i\right)\right] = \alpha + \beta_0(t_i - \bar{t}) + \beta_1(t_i - \tau_1)^+.$$

The model with no breakpoint is named as M_0, one breakpoint as M_1 and so on. There will be M_{k+1} nested models over the model space in total depending upon the number of breakpoints.

In the proposed model given in (1), α, and β_0 represent the common parameters where as $\beta_j^{'}s$ are non-common parameters that are model-specific. β_0 together with $\beta_j^{'}s$ gives the slope for the different models with at least one change point. To give the same meaning across different models for all common parameters, Martinez-Beneito et al. (2011) proposed an alternative parametrization imposing different conditions. This work is motivated by their work and follows the same parametrization.

The purpose of this study is to fit the joinpoint regression model for the childhood brain and other CNS cancer mortality counts. This model is based on its probabilistic framework that provides a reliable estimates of annual mortality trend. Because the behavior of the mortality count data in the population is a rare event, characterized by Poisson distribution $(Y_i, Poi(\lambda_i, i = 1, 2, \cdots, n))$, it is modeled using natural log link function. Hence, the model in the equation (1) becomes

$$\log(\lambda_i) = \log(n_i) + \alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \delta_j \beta_j B_{\tau_j}(t_i) \qquad (2)$$

where n_i is the total number of population at time t_i, $B_{\tau_j}(t)$ is the piecewise linear function reparametrized as in Martinez-Beneito et al. (2011), called as break-point centered at τ_j, and $\delta_j, j = 1, 2, \cdots, k$ are binary indicator variables for the inclusion or exclusion of the change points in the model i.e.

$$\delta_j = \begin{cases} 1 & \text{for each breakpoint} \\ 0 & \text{otherwise} \end{cases}$$

The above equation (2) leads to the following estimated rate:

$$E(r_i) = exp(\alpha + \beta_0(t_i - \bar{t}) + \sum_{j=1}^{k} \delta_j \beta_j B_{\tau_j}(t_i)). \quad (3)$$

The annual percentage change(APC) is used to characterized the trends or the change in rates over time. APC from i^{th} year to $(i+1)^{th}$ year is given as

$$APC_i = \frac{r_{i+1} - r_i}{r_i} \times 100.$$

Because the model can choose an infinite number of breakpoints, it is necessary to impose some restrictions on the position of the change points in the model. This is done by assigning minimum gap of two years between two joinpoints starting after the first years and ending before the last two years.

The aim is to find the trend that describes the behavior of the data well. This will be carried out by detecting the points and their locations where the significant changes occur within the data range. Finding such locations in this model selection problem is carried out by using Bayes Factor in which data updates the prior odds to yield posterior odds. Bayes Factor summarizes the relative support for one model versus another for all competing models by selecting a model with highest posterior probability. Therefore, the posterior probability of each model is calculated and the one with highest posterior probability is selected as the best model.

The specification of priors plays a major role in Bayesian model selection problem. In an objective Bayes solution to the model selection problem, the nature of the posterior distributions depends upon the selection of priors and is very sensitive if there are non-common parameters in the models as explained in Berger and Pericchi (2001), and Bayarri and García-Donato (2008). Furthermore, the choice of improper or vague priors would lead to arbitrary Bayes Factor and make the result computationally challenging (see Charlin et al., 1992; Martinez-Beneito et al., 2011). For the commmon parameters α, and β_0, choose flat priors i.e. $\pi(\alpha, \beta_0) \propto 1$. For non-common parameters, the divergence-based (DB) priors introduced in Bayarri, et al. (2008) as a generalization of the ideas discussed in Zellner and Siow (1980), Jeffreys (1961), and Zellner (1984) and implemented in Martinez-Beneito et al. (2011) is considered. The parameter space for τ is bounded, and hence the default prior $\pi(\tau) \propto 1$ was chosen. Based on the nature of

δ, it is reasonable to choose independent Bernoulli priors with a probability of success p with hyper priors for p being $Beta(\frac{1}{2}, \frac{k-1}{2})$ where k is the number of joinpoints in the model.

In Bayesian paradigm, finding a good candidate model from a set of nested models can be computationally intensive. The high dimensionality of the integrals makes the model selection procedure even more complex. In choosing priors, the distribution of the posterior probability is not analytically tractable, thus Gibbs variable selection approach in WinBUGS software is used to select the best model with significantly minimum number of joinpoints that describes the trend. The process is carried out in such a way that if one more joinpoint is added in the model, the model becomes insignificant.

Results

To apply the model discussed, annually observed mortality counts for childhood brain and other CNS cancers from the Surveillance Epidemiology and End Results (SEER) database of National Cancer Institute (NCI) from 1969-2009 were used. The data were extracted from publicly used database of the SEER program 7.1.0 with the adjustments of Katrina/Rita population.

The joinpoint model is fitted using WinBUGS software. The model is described by four unknown joinpoints ($k = 4$) to identify the time where changes in the slope of child brain cancer mortality trend occurs. Two parallel chains using different initial values are implemented. Each chain is run for 150,000 iterations giving 50,000 iterations as burn-in period. The posterior inferences is based on 100,000 iterations for each chain combining total of 200,000 iterations for each of the parameters. The posterior summaries for the parameters are given in Table 1. Out of competing five nested models, the model selection procedure selected the model with one joinpoint as given in Figure 1 (left). For the selected model with one joinpoint, the posterior distribution of each of the parameters is observed by monitoring the trace, iterations, Monte Carlo errors, standard deviations, and density curves. The trace for each of the parameters satisfy the convergence criteria. Also, the Monte Carlo errors are within 0.1% of the posterior standard deviations.

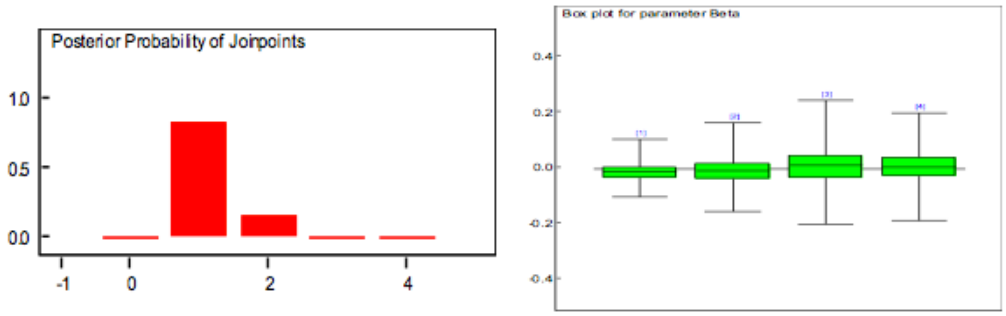

Figure 1: Posterior distribution of the number of joinpoints in child brain cancer mortality trend in United States (left), Box plot for parameters of joinpoints (right).

Table 1: Parameter estimates

node	mean	sd	MC error	2.50%	median	97.50%	start	sample
alpha	-11.76	0.006448	3.35E-05	-11.77	-11.76	-11.75	50000	200002
beta0	-0.01176	5.33E-04	2.79E-06	-0.01281	-0.01176	-0.01071	50000	200002
beta[1]	-0.0176	0.05287	7.68E-04	-0.09726	-0.02668	0.09301	50000	200002
beta[2]	-0.01679	0.09534	0.001723	-0.1736	-0.02925	0.1602	50000	200002
beta[3]	-0.00151	0.1265	0.001355	-0.218	-0.00167	0.2119	50000	200002
beta[4]	-7.90E-04	0.1114	0.001049	-0.1963	-1.52E-04	0.1938	50000	200002
delta[1]	0.5254	0.4994	0.01384	0	1	1	50000	200002
delta[2]	0.4684	0.499	0.01359	0	0	1	50000	200002
delta[3]	0.1156	0.3197	0.005156	0	0	1	50000	200002
delta[4]	0.05771	0.2332	0.001234	0	0	1	50000	200002

As depicted in the graph given in Figure 1 (left), the probability of the posterior distribution for one joinpoint is about 80%. The probability of the posterior distribution for no joinpoint is very low indicating that the linear trend is not a choice. Similary, the probability of posterior distribution does not support two, three, and four joinpoints as well. The boxplot for the parameters β_j, $j = 1, 2, 3, 4$ associated with change points is plotted in Figure 1 (right). Posterior means and 95% credible intervals of β_j's suggest that their posterior distributions are not discriminable. This indicates that no more than one joinpoint

is required and if more joinpoints are added, the model is not statistically significant.

The estimated rates for each year from 1969-2009 are obtained by averaging the estimates of joinpoint and other parameters in the model at every step of MCMC. The graph for the estimated rate and its prediction is given in Figure 2. The solid curve represents the estimated trend line for annual mortality rate whereas the dashed lines represent its 95% pointwise credible interval. The observed death rates are represented by unfilled circles. The extended graph beyond dashed vertical line represents the prediction of rate from 2010 to 2012. The predicted rates are obtained by averaging the joinpoint curve at every steps of the MCMC from the posterior predictive distribution.

The graph shows that the childhood cancer mortality rates declined faster from 1969 to 1978 compared to the rest of the time interval in a decreasing fashion. The overall mortality rate decreased from 1.056 to 0.63 per 100,000 by 2009 and is predicted to decrease continuously.

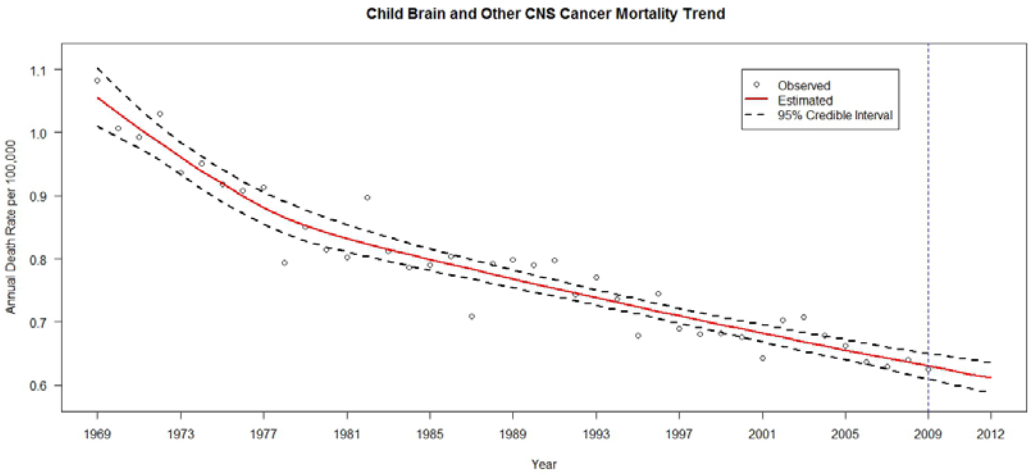

Figure 2: Estimated time trend for the annual observed mortality rate per 100,000 children

For the same data, the joinpoint regression model is fitted using the joinpoint software of NCI. The model was fitted with the assumption of Poisson variance using crude death rate with an autocorrelated errors based on the data. Here, the heteroscedasticity is conducted by joinpoint using weighted least square. Grid search method is used to select the joinpoint model with grid size of 2 years

leaving two years at the two ends of the data values. This was done to exactly match the condition imposed for identifiability problem in the Bayesian joinpoint model. The model selection method was performed using permutation test for four joinpoints which performs multiple tests to select the number of joinpoints using the Bonferroni correction at 0.05 overall significance level for multiple testing. The output is as shown in Figure 3.

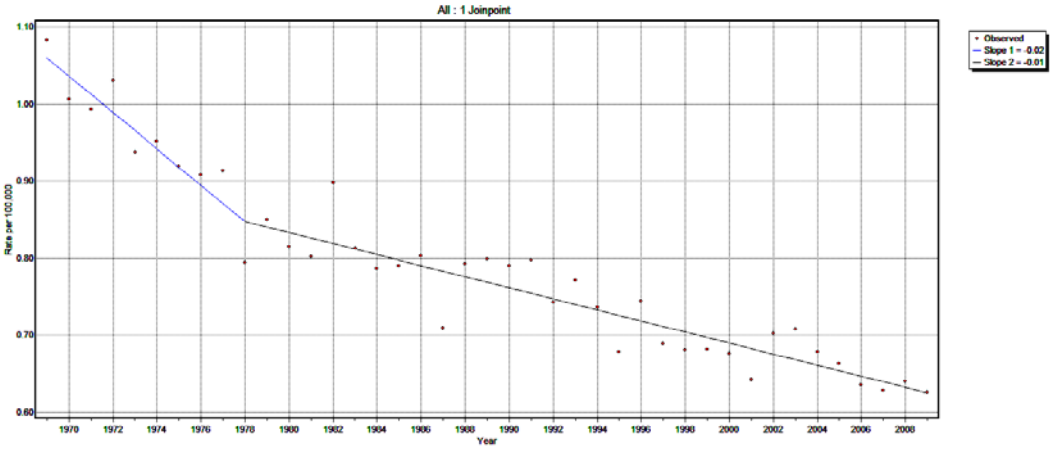

Figure 3: Mortality rates of child brain cancer(1969-2009) using the joinpoint software of NCI.

The solid line represents the estimated mortality rates obtained by using the joinpoint software of NCI. The graph shows that there is one joinpoint observed exactly at 1978. The trend line is piecewise linear indicating that the slopes of the rate curve before and after joinpoint are constant. It is not the case for the applied Bayesian joinpoint model as it gives the slope of the rate curve at any point. Also, the location of change point is discrete and occurs exactly at the whole number year in case of the regression trend given by joinpoint software whereas the location of change point is continuous in this case and can occur in between the years. Another difference is that the trend obtained from joinpoint software is descriptive but the regression trend obtained can give the insights for the mortality trend in future with credible bands.

The graph in Figure 4 gives the average rate of change in mortality rate per year from 1969 to 2009 and its predictions up to 2011. APC is approximately -2.31 for the first three years and increases from -2.29 in 1973 to -1.12 in 1980.

After 1980, APC looks almost contant with a fluctuation of 0.01 to 0.02 over the entire range. It means that the average rate of change per year in childhood brain cancer mortality rate has not been changed in recent years and is predicted to remain almost the same in the consequent years.

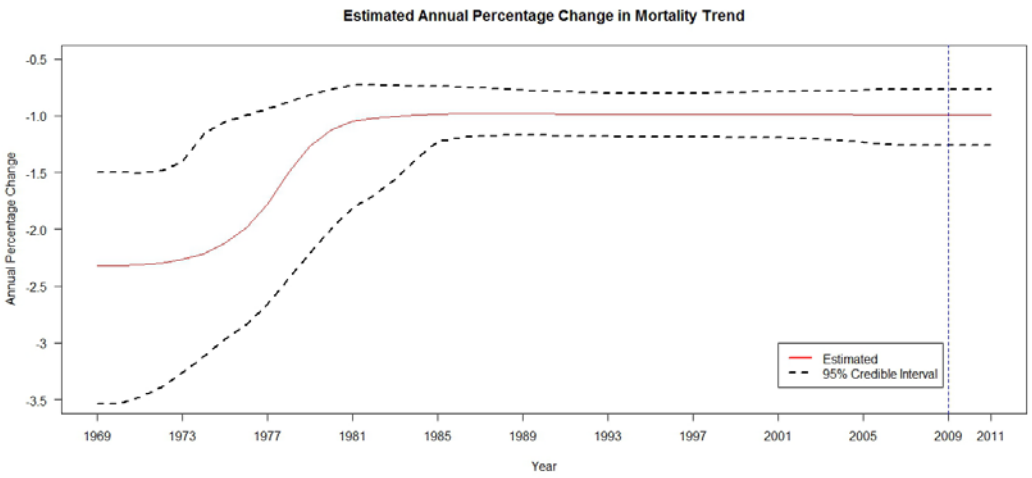

Figure 4: Estimated Annual Percentage Change in child brain cancer rates over time per 100,000 children

---

To check the validity, goodness of fit, and the assumptions of the proposed model, different model validation techniques discussed in literature are performed. The residual analysis is performed to check the robustness and fit of the developed model. The mean and standard deviation of the standardized residual are 0.000527 and 0.927 respectively. This indicates that the developed model fits the observed data well. The Chi-square statistics for the observed mortality data as well as for the predicated data in each iteration of MCMC are calculated.The difference between two statistics is monitored and their corresponding posterior p-value is obtained. The p-value based on the difference of Chi-squares obtained as a posterior mean using WinBUGS is 0.5513. The large p-value shows that the observed statistic is close from what is expected under the assumed model. Also, the observed mortality counts fall not only inside the 95% posterior intervals of replicated data but also close to their mean values indicating that the assumptions of Poisson distribution is valid.

Conclusion

This study applied newly developed Bayesian joinpoint regression model to uncover the patterns of childhood brain cancer mortality that provides an important information pertaining further study in the cases and control of the disease. Although, different studies have shown that the childhood cancer mortality rates continue to decline dramatically by more than 50% in the past two decades (Ries, et al., 2007; Kohler, et al., 2011) in the United States, only few studies have considered the probability distribution of the observed counts as Poisson and the location of the change points continuous in time. The application discussed here based on these probabilistic assumptions. The trend is obtained such that it describes the behavior of the observed data very well and gives the best possible short term predictions. The temporal trend provides the different slopes of the rate curve at each point of time. In contrast, the joinpoint software of NCI gives the same slope at each year between two change points. Also, it was possible to obtain the more accurate annual percentage change (APC) and it is observed that the APC is almost constant from 1981 and is predicted to remain constant. SEER routinely collects the data covering 28% of the US population and there is a three years lag in time to collect and process the data. In this scenario, predictions in the temporal trend and APC are very helpful to evaluate the effectiveness of the current status of the disease and play an important role to make evidence based policy. This improvement over the existing methods allow observation of the real progress being made in childhood brain cancer.

This work may be extended to study the influence in the mean of the outcome by incorporating covariates in the model. But the addition of covariates increases the complexity of the model. The Bayes Factors are sensitive to the prior specifications, and therefore further study is needed in selecting the objective priors by exploring different objective model selection criteria for priors that can deal with model uncertainty. Moreover, age standardized rates in this methodology could be a future extension. Also, studying incidence and mortality rates at the same time will depict the clear picture of real improvements being made in cancer research.

Acknowledgements

The authors wish to thank the University of Tampa Dana Foundation Grant.

References

Bayarri, M. J., & García-Donato, G. (2008). Generalization of Jeffreys' divergence based priors for Bayesian hypothesis testing. *Journal of the Royal Statistics Society; Series B (Statistical Methodology, 70*(5)**,** 981-1003.

Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: introduction and comparison (with discussion). *Model Selection, 38*, 135-207. Institute of Mathematical Statistics.

Carlin, B. P., Gelfand, A. E., & Smith, A. F. M. (1992). Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics, 41*(2), 389-405.

Ghosh, P., Basu, S., & Tiwari, R. C. (2009). Bayesian analysis of cancer rates from SEER program using parametric and semiparametric joinpoint regression models. *Journal of the American Statistical Association, 104*(486), 439-452.

Ghosh, K., & Tiwar, R. C. (2007). Prediction of U.S. cancer mortality counts using semiparametric Bayesian techniques. *Journal of the American Statistical Association, 102*(477), 7-15.

Jeffreys, H. (1961). Theorey of Probability. *London: Oxford University Press*, 3rd edition.

Kim, H., Fay, M. P., Feuer, E. J., & Midthune, D. N. (2000). Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine, 19*(3), 335-351.

Kim, H. J., Fay, M., Yu, B., Barrett, M.J., & Feuer, E.J. (2004). Comparability of segmented line regression models. *Biometrics, 60*(4), 1005-1014.

Kleihues, P., Burgers, P. C., Scheithauer, B. W. , et al. (1993). *World health organization histological typing of tumors of the central nervous system*. New York: Springer-Verlag.

Kohler, B. A., Ward, E., McCarthy, B. J., et al. (2011). Annual report to the nation on the status of cancer, 1975-2007, featuring tumors of the brain and other nervous system. *Journal of National Cancer Institute, 103*(9), 714-736.

Levy, A. S. (2005). Brain tumors in children: evaluation and management. *Current Problems in Pediatric and Adolescent Health Care, 35*, 230-244.

Martinez-Beneito, M. A., Garcia-Donato, G., Salmeron, D. A. (2011). Bayesian joinpoint regression model with an unknown number of break-points. *Annals of Applied Statistics, 5*(3), 2150-2168.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New York: Wiley Publication.

Peterson, K. M., Shao, C., McCarter, R., MacDonald, T., & Byrne, J. (2006). An analysis of SEER data of increasing risk of secondary malignant neoplasmsamong long-term survivors of childhood brain tumors. *Pediatric Blood Cancer*, *47*(1), 83-88.

Pollack, I. F. (1994). Brain tumors in children. *The New England Journal of Medicine*, *331*(22), 1500-1507.

Pollack, I. F. (1999). Padiatric brain tumors. *Seminars in Surgical Oncology*, *16*(2), 73-90.

Ries, L., Melbert, D., & Krapcho, M. (2007). *SEER Cancer Statistics Review, 1975-2004*. National Cancer Institute.

Surveillance Epidemiology and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Mortality - All COD, Aggregated With State, Total U.S. (1969-2009) <Katrina/Rita Population Adjustment>.

Serveillance Research, National Cancer Institute, Joinpoint Regression program (serveillance.cancer.gov/joinpoint).

Tiwari, R. C., Cronin, K. C., Davis, W., Feuer, E. J., Yu, B., & Chib, S. (2005). Bayesian model selection for joinpoint regression with application to age-adjusted cancer rates. *Applied Statistics*, *54*(5), 919-939.

Ullrich, N. J., & Pomeroy, S. L. (2003). Pediatric brain tumors. *Neurologic Clinics*, *21*(4), 897-913.

Zellner, A. (1984). Posterior odds ratios for regression hypothesis: general considerations and some specific results. *Basic Issues in Econometrics*, 275-305. Chicago: University of Chicago Press.

Zelner, A., & Siow A. (1980). Posterior odds ratio for selected regression hypotheses. *In Bayesian Statistics 1* (J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith, eds.), *31*(1), 585-603. Valencia: University Press.